Molecular Diversity (2021) 25:1553–1568
https://doi.org/10.1007/s11030-021-10245-z

ORIGINAL ARTICLE

Machine learning models to select potential inhibitors of acetylcholinesterase activity from SistematX: a natural products database

Chonny Herrera-Acevedo[1,4] · Camilo Perdomo-Madrigal[2] · Kenyi Herrera-Acevedo[3] · Ericsson Coy-Barrera[4] · Luciana Scotti[1] · Marcus Tullius Scotti[1]

Received: 31 March 2021 / Accepted: 3 June 2021 / Published online: 16 June 2021

Abstract

Alzheimer's disease is the most common form of dementia, representing 60–70% of dementia cases. The enzyme acetylcholinesterase (AChE) cleaves the ester bonds in acetylcholine and plays an important role in the termination of acetylcholine activity at cholinergic synapses in various regions of the nervous system. The inhibition of acetylcholinesterase is frequently used to treat Alzheimer's disease. In this study, a merged BindingDB and ChEMBL dataset containing molecules with reported half-maximal inhibitory concentration (IC_{50}) values for AChE (7032 molecules) was used to build machine learning classification models for selecting potential AChE inhibitors from the SistematX dataset (8593 secondary metabolites). A total of seven fivefold models with accuracy above 80% after cross-validation were obtained using three types of molecular descriptors (VolSurf, DRAGON 5.0, and bit-based fingerprints). A total of 521 secondary metabolites (6.1%) were classified as active in this stage. Subsequently, virtual screening was performed, and 25 secondary metabolites were identified as potential inhibitors of AChE. Separately, the crystal structure of AChE in complex with (–)-galantamine was used to perform molecular docking calculations with the entire SistematX dataset. Consensus analysis of both methodologies was performed. Only eight structures achieved combined probability values above 0.5. Finally, two sesquiterpene lactones, structures **15** and **24,** were predicted to be able to cross the blood–brain barrier, which was confirmed in the VolSurf+ quantitative model, revealing these two structures as the most promising secondary metabolites for AChE inhibition among the 8593 molecules tested.

✉ Marcus Tullius Scotti
mtscotti@gmail.com

1 Post-Graduate Program in Natural and Synthetic Bioactive Products, Federal University of Paraíba, João Pessoa, PB 58051-900, Brazil

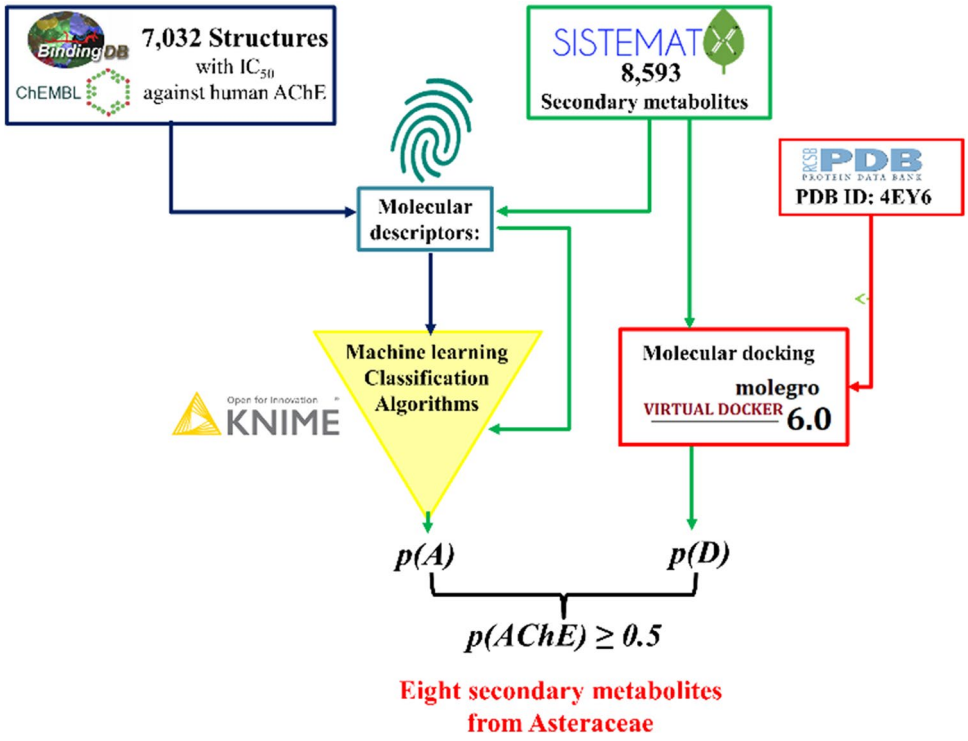
2 School of Science, Universidad de Ciencias Aplicadas y Ambientales, Calle 222 # 55–37, Bogotá D.C., Colombia

3 Departamento de Química, Facultad de Ciencias, Universidad Nacional de Colombia, Carrera 45 # 26- 85, Bogotá D.C., Colombia

4 Bioorganic Chemistry Laboratory, Facultad de Ciencias Básicas y Aplicadas, Universidad Militar Nueva Granada, 250247 Cajicá, Colombia

Graphic abstract

A consensus analysis of classification models and molecular docking calculations identified four potential inhibitorsof acetylcholinesterase from the SistematX dataset (8593 structures).



Keywords Natural products · Machine learning · Acetylcholinesterase · Ligand-based virtual screening · Machine learning · SistematX database

Introduction

Acetylcholinesterase (AChE, E.C. 3.1.1.7) is the enzyme responsible for the cleavage of ester bonds in acetylcholine and is one of the most active enzymes in the human body, associated with a turnover rate of $1.5 \times 10\ 4\ s^{-1}$ [1]. AChE plays an important role in the termination of the acetylcholine activity at the cholinergic synapses in various regions of the nervous system [2, 3]. This enzyme is a homomeric protein formed by three catalytic subunits and is typically bound to a collagen-like subunit in the neuromuscular junction [4, 5].

Acetylcholinesterase inhibition is frequently used as a therapeutic strategy for several diseases in humans, especially Alzheimer's disease [6, 7], which is the most common neurodegenerative disease related to aging and results in the loss of mental acuity, functional declines, and disrupts learning abilities [8]. Among the most popular AChE inhibitors, galantamine, donepezil, rivastigmine, and tacrine act as reversible inhibitors and are commonly used for therapeutic purposes [7, 9]. Irreversible AChE inhibitors, such as organophosphorus compounds, exert toxic effects on the organism[7, 10]. Approved drugs typically demonstrate good effectiveness for controlling disease symptoms, but some are not specific to AChE, exerting inhibitory effects against other cholinesterases. All currently approved AChE inhibitors are associated with secondary effects, including dizziness, anorexia, nausea, abdominal pain, and others [7].

To identify more specific and effective drugs against Alzheimer's disease, great interest exists in the identification of chemotherapeutic molecules found in natural sources, such as plants and fungi [11]. These compounds exhibit great structural diversity and complexity, which is difficult to reproduce using conventional synthetic methods. The extended use of natural remedies in many traditional medicinal treatments among various populations demonstrates their acceptable effectiveness against a variety of diseases [12, 13]. Huge quantities of information are available as a result of global research efforts to identify these metabolites, which have been stored and shared with the scientific community through multiple databases, including Dictionary of Natural Products, NUBBEdb, SuperNatural II, KNApSAcK Family,

Collection of Open Natural Products database (COCONUT), and others [14–17].

SistematX (http://sistematx.ufpb.br) is an emergent open-access database for secondary metabolites, which can be used by any research group. SistematX was developed by the cheminformatics laboratory of the Federal University of Paraiba and contains a wealth of useful information regarding secondary metabolites, highlighting the exact species and locations from which compounds were isolated [18]. Currently, SistematX contains approximately 9,000 unique secondary metabolites, including approximately 20,000 unique botanical occurrences from five botanical families: Asteraceae, Apocynaceae, Annonaceae, Lamiaceae, and Solanaceae.

Databases have become key tools in chemoinformatic investigations and are indispensable for computer-aided drug design (CADD) studies. The biological activity information available in these databases allows for the calculations of molecular descriptors of the identified metabolites, and many accessible CADD tools exist for designing various models to identify and predict biological activity and propose potential modifications of molecular motifs that could potentially increase biological activity. Quantitative structure–activity relationship (QSAR) and machine learning models associated with molecular docking methodologies are typically designed for these purposes [10, 19].

This study applied a combined approach in which machine learning classification models and molecular docking calculations were used to identify secondary metabolites from the SistematX dataset (8,593 structures) with potential inhibitory activity against AChE. Initially, machine learning classification models were built using three types of classificatory algorithms and molecular descriptors obtained from a dataset consisting of 7,032 molecules with reported *in vitro* AChE inhibitory activity. In parallel, molecular docking calculations were performed using the crystal structure of recombinant human AChE in complex with (–)-galantamine (Protein Data Bank [PDB] ID: 4EY6). Finally, a consensus analysis of the two methodologies was performed, using the probability values calculated throughout the study to select those molecules with potential activity against AChE.

Material and method

Database

From the BindingDB [20] and ChEMBL databases (https://www.ebi.ac.uk/chembl/), we selected a diverse set of structures that were initially classified according to their calculated activity against human AChE, including 6766 structures from BindingDB and 7479 structures from ChEMBL 202. These compounds were classified according to their pIC_{50} values (–log[half-maximal inhibitory concentration (IC_{50})] in mol/L); therefore, we stratified these structures into active ($pIC_{50} \geq 6.0$) and inactive ($pIC_{50} < 6.0$). The activity cutoff was selected based on the reported pIC_{50} value of galantamine (5.97 ± 0.03), which was used as a control in this study and represents one of the primary drugs used in the treatment of Alzheimer's disease [21, 22]. Most reversible inhibitors used for therapeutic purposes exhibit either competitive or non-competitive AChE inhibitory interactions. Due to the variability of experimental protocols used to obtain the data presented in both databases, a qualitative pattern was used to partially minimize the differences in activity values associated with different experimental protocols and strains [23]. The IC_{50} values represent the concentration necessary to inhibit 50% of AChE activity. Data curation was performed on the datasets according to the procedures suggested in the literature [24, 25]. Standardizer software [Jchem, version 20.19.0.708 (2020), calculation module developed by ChemAxon, http://www.chemaxon.com/] was used to canonize all simplified molecular-input line-entry system (SMILES) codes. After duplicate structures were removed, those with higher pIC_{50} values were eliminated. The use of only those compounds with lower activity values facilitated the generation of more restrictive models.

After dataset curation, the datasets were combined, resulting in a dataset containing 7,032 unique structures with reported AChE activity (3010 active and 4022 inactive). For all structures, the SMILES codes were used as the input data in Marvin [ChemAxon, version 20.19.0.708 (2020), calculation module developed by ChemAxon, http://www.chemaxon.com/]. We used Standardizer software [Jchem, version 20.19.0.708 (2020), calculation module developed by ChemAxon, http://www.chemaxon.com/] and ChemAxon to canonize the structures, add hydrogens, perform aromatic form conversions, and clean the molecular graphs in three dimensions. This software was used to generate and optimize conformers for the initial structure (represented by the root node in the tree). Those molecules that presented structural problems during the three-dimensional (3D) structure generation were manually corrected using Marvin sketch [23].

The applicability domain (APD), based on Euclidean distances, was used to identify those compounds in the test set for which predictions may be unreliable. Compounds were considered unreliable if they had APD values higher than

$d + Z\sigma$, where d was the average Euclidian distance, and σ was the standard deviation of the set of samples in the training set with lower-than-average Euclidian distance values relative to all samples in the training set. The parameter Z is an empirical cutoff value, and 0.5 was used as the default value [26].

SistematX secondary metabolites

The entire dataset of SistematX database [18] (8,653 secondary metabolites) was obtained in. comma-separated value (CSV) format. All SMILES codes were canonized using Standardizer software [Jchem, version 20.19.0.708 (2020), calculation module developed by ChemAxon, http://www.chemaxon.com/]. After duplicate structures were removed, 8,593 unique structures were identified**.** ChemAxon Standardizer also was used to generate 3D structures using the following options: add hydrogens, perform aromatic form conversions, and clean molecular graphs in three dimensions.

Molecular descriptors

Volsurf+ descriptors

The 3D structures of the identified molecules, in special data file (SDF) format, were used as input data in VolSurf+ v. 1.0.7 and were subjected to molecular interaction fields (MIFs) to generate descriptors using the following probes: N1 (amide nitrogen–hydrogen-bond donor probe), O (carbonyl oxygen–hydrogen-bond acceptor probe), OH_2 (water probe) and DRY (hydrophobic probe). Additional non-MIF-derived descriptors were generated, resulting in a total of 128 descriptors. One of the main advantages of using VolSurf+ descriptors is the relatively low influence of conformational sampling and averaging on these descriptors [27, 28].

DRAGON descriptors

DRAGON 5.0 computer software [29, 30] was employed to calculate 1664 molecular descriptors. For all descriptors, constant variables were excluded, and only those that presented different values were retained. For the remaining descriptors, pairwise correlation ($r < 0.99$) analysis was performed to exclude those that were highly correlated. Thus, the number of DRAGON descriptors used in our calculations was reduced to 1437 [31].

RDK bit-based fingerprints

Generates hashed bit-based fingerprints for an input of the 3D structures of the identified molecules in SDF format. A total of 1,024 bit-based fingerprints were calculated using the circular fingerprint based on the Morgan algorithm and connectivity invariants (ECFP-like) in KNIME 4.3.2. through "RDK Fingerprint" and the "Expand bit vector" nodes [32].

Machine learning classification models

KNIME 4.3.2 software (KNIME 4.3.2 the Konstanz Information Miner Copyright, 2003–2014, www.knime.org) was used to perform all analyses. Initially, molecular descriptors calculated in the VolSurf+, DRAGON, and bit-based fingerprint methods were imported, separately, in CSV format. The "Partitioning" node in the stratified sampling option was used to classify 90% of the initial dataset as the training set, and the remaining 10% was used as the test set. The "X-Partitioning" node in the stratified sampling option was used to divide the dataset five times into a modeling set (80%–20%), to perform a "fivefold cross-validation" procedure using WEKA nodes. The models were generated by employing three PMML algorithms with the following specific parameters: a) Decision Tree Learner (Gini index, no pruning, number of threads: 8, and minimal number of records per node: 2); b) Gradient Boosted Tree Learner (tree depth: 4, number of models: 100, and learning rate: 0.1); and c) Support Vector Machine, Polynomial (Gamma 1.0, Bias 1.0, Power 1.0, and overlapping penalty 1.0). From the confusion matrix, the internal and external performances of the selected models were analyzed, using the following parameters: sensitivity (true-positive rate), specificity (true-negative rate), and accuracy (overall predictability). Because an imbalanced dataset is generated, the Synthetic Minority Oversampling TEchnique (SMOTE) was used to perform the redistribution of the imbalanced dataset. SMOTE is a technique for generating new synthetic instances to rebalance class distributions. SMOTE creates extra synthetic instances by pairing each positive instance with its nearest neighbor and generates new instances along a line segment between the pair of positive instances [33]. This procedure was performed using the five nearest neighbors and an oversample of two. In addition, to describe the true performance of the model with more clarity than can be obtained from accuracy alone, the receiver operating characteristic (ROC) curve was employed, using a "ROC curve" node, which uses the sensitivity and specificity parameters. The plotted ROC curve shows the true-positive (active) rate versus the false-positive rate (1-specificity) [34]. In this representation, when a variable of interest cannot be distinguished between the two groups, the area under the ROC curve (AUC) value is 0.5, whereas a perfect separation between the values of the two groups, with no distribution overlap, results in an AUC value of 1. Matthew's correlation coefficient (MCC) was also calculated, for which a value of 1 represents a perfect prediction, a value of 0 represents a random prediction, and a value of −1 represents total disagreement between the prediction and the observation [35].

Molecular docking calculations

For molecular docking calculations, the structure of the recombinant human AChE (PDB ID: 4EY6) in complex with (–)-galantamine (PDB ID: GLN), a reversible, competitive, tertiary alkaloid AChE inhibitor, was downloaded from PDB [3, 36]. This molecular docking protocol was designed to identify exclusive secondary metabolites that use the same inhibition mechanism that galantamine uses against AChE. Using Molegro 6.0.1 software, both rigid and flexible approaches were performed. All water compounds were deleted from the enzyme structure, and the enzyme/compound structures were prepared using the same default parameter settings in the same software package (score function: MolDock Score; ligand evaluation: Internal ES, Internal HBond, Sp2–Sp2 Torsions, all checked; number of runs: 10 runs; algorithm: MolDock SE; maximum interactions: 1500; max. population size: 50; max. steps: 300; neighbor distance factor: 1.00; max. number of poses returned: 5). The rigid docking procedure was performed using a grid with an 18-Å radius, a 0.30-Å resolution, and a binding site center of X: −9.94, Y: −43.49, and Z: 30.29. For flexible molecular docking calculations, the residues within a distance of 5 Å from the AChE pocket were set as flexible, totaling 28 residues.

Drug-like properties of the potential inhibitors against AChE

The absorption, distribution, metabolism, and excretion (ADME) parameters were calculated for all secondary metabolites classified as actives against AChE in the consensus analysis using the SwissADME server [37], an open-access web tool (http://www.swissadme.ch). Drug toxicity predictions were performed in OSIRIS Data Warrior v.5.2.1, based on the following parameters: mutagenicity, tumorigenicity, reproductive effects, and irritability [38]. The VolSurf+ model for blood–brain barrier (BBB) permeation is a quantitative model containing approximately 500 related, but chemically diverse compounds that were extracted from the literature and an in-house database, which are defined as either brain-penetrating (Exp. LgBB > 0.5), have moderate permeation (LgBB between 0 and 0.5), possess little ability to cross the BBB (Exp. LgBB greater than −0.3), or demonstrate very little permeation (LgBB less than −0.3) [27, 28, 39].

Results and discussion

Machine learning classification models were built based on structures with previously demonstrated inhibitory activity against human AChE that were registered in the ChEMBL database (https://www.ebi.ac.uk/chembl/) and BindingDB. All procedures were performed in accordance with good practices established for QSAR studies, and the entire dataset was curated using the procedures suggested in the literature [24, 25].

A total of 7032 unique structures were analyzed in this study following curation (*AChE dataset*). These structures were used to develop models using three types of molecular descriptors (Volsurf [27, 28], DRAGON 5.0 [29, 30] and Bit-based fingerprints) and three algorithms (Decision Tree Learner, Gradient Boosted Tree Learner, and Support Vector Machine) to identify additional human AChE inhibitors from the SistematX dataset containing 8,593 secondary metabolites (*SistematX dataset*). SistematX is a web tool developed by the Laboratory of Cheminformatics of the Federal University of Paraiba that contains a wealth of useful information for the scientific community regarding natural products, including the locations of those species from which various compounds were isolated [18].

Subsequently, for both the 7,032-structure *AChE dataset* and the 8,594-structure SistematX secondary metabolite dataset, VolSurf+ [27, 28] , DRAGON 5.0 (1664) [29, 30] and bit-based fingerprint (1024) descriptors were calculated using KNIME 4.3.2. software (KNIME 4.3.2. the Konstanz Information Miner Copyright, 2003–2014, https://www.knime.org) [32]. The classification models were generated using three algorithms (Decision Tree Learner, Gradient Boosted Tree Learner, and Support Vector Machine) and validated through fivefold cross-validation to assess their abilities to determine activity probabilities for the entire SistematX dataset (Fig. 1).

Additionally, molecular docking calculations were performed using the structure of recombinant human AChE (PDB ID: 4EY6) in complex with (–)-galantamine (PDB ID: GLN) [3] in Molegro virtual docker 6.0 software. A consensus analysis was performed to obtain probability values for each structure in the SistematX dataset, based on the docking results obtained from the classification models, to identify potentially reversible AChE inhibitors. In addition, a docking validation was performed using the AChE dataset (both active and inactive structures), using docking calculations for each of the identified 7032 structures. A ROC curve was built to redistribute the imbalanced dataset using SMOTE [33].

Machine learning classification models

Initially, three types of molecular descriptors were calculated for the entire 7,032-molecule AChE dataset: VolSurf+ descriptors (128 values), DRAGON 5.0 descriptors (1,664 values), and bit-based fingerprint descriptors (1,024). The descriptors were classified using a binary classification system (as active or inactive). Those molecules with $pIC_{50} \geq$

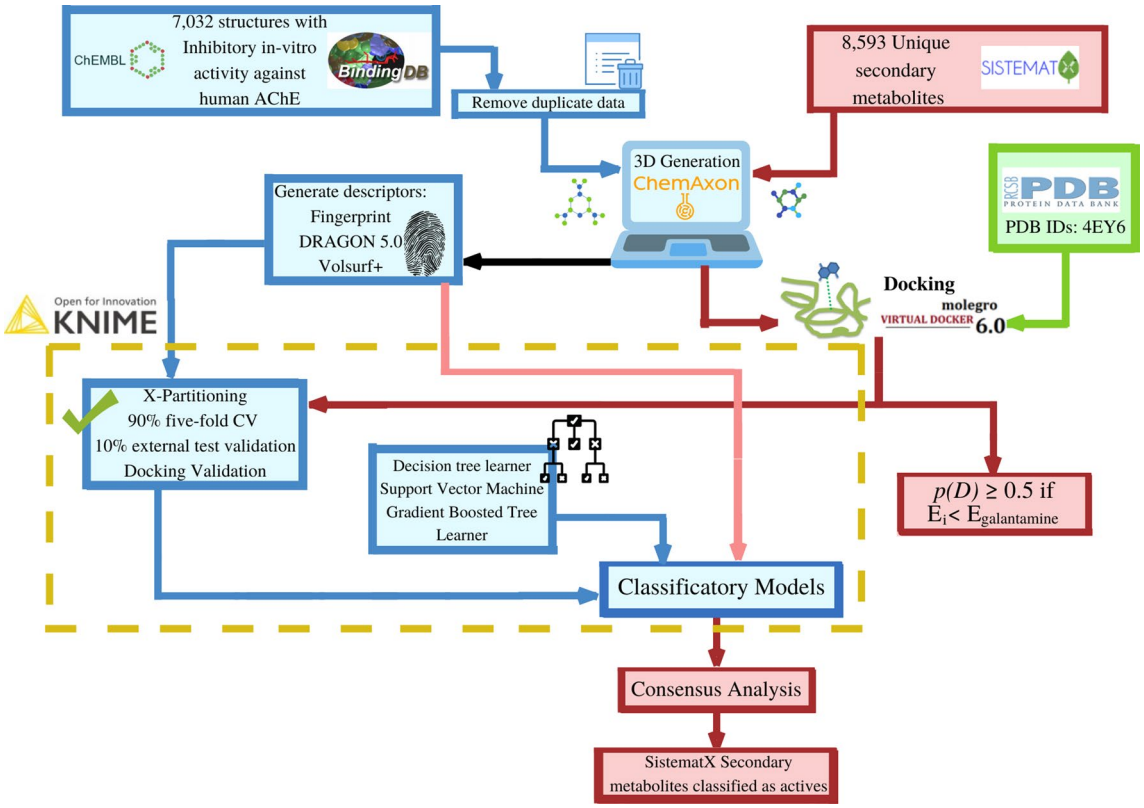


Fig 1 The virtual screening methodology used in this study. Solid blue lines represent the AChE dataset used to generate and validate the machine learning classification models using three types of descriptors: VolSurf+, DRAGON 5.0, and bit-based fingerprints. The red lines represent the procedures used to evaluate 8,593 secondary metabolites obtained from SistematX. The black line represents both datasets (AChE and SistematX). The green line represents the recombinant human AChE (PDB ID: 4EY6) in complex with (–)—galantamine (PDB ID: GLN), which was extracted from the Research Collaboratory for Structural Bioinformatics (RCSB) Protein Data Bank (PDB). The yellow dash-dot border delimits the process performed in KNIME software

6.0 (1 μM or 1×10^{-6} M) were classified as active, based on the reported pIC_{50} value of galantamine (5.97 ± 0.03), which is one of the primary AChE inhibitors approved for Alzheimer's disease treatment and served as the positive control for this study [21, 22]..

The three types of molecular descriptors were input into the KNIME program in CSV format, resulting in the generation of seven classification models [32] employing three algorithms (Decision Tree Learner, Gradient Boosted Tree Learner, and Support Vector Machine). Several models were evaluated to minimize the false-positive rates of the models. For each algorithm, the following parameters were selected: a) *Decision Tree Learner* (Gini index, no pruning, number of threads: 8, and minimal number of records per node: 2*); b) Gradient Boosted Tree Learner* (tree depth: 4, number of models: 100, and learning rate: 0.1); and *c) Support Vector Machine,* Polynomial (Gamma 1.0, Bias 1.0, Power 1.0, and overlapping penalty 1.0)**.**

Structures with pIC_{50} values between 5.8 and 6.0 (range of 0.2 units) were excluded to avoid edge effects and improve the predictive capacity of the models by minimizing potential activity differences due to errors and the use of different experimental protocols [23]. The performances of the classification models were verified through fivefold cross-validation, splitting the dataset five times into a modeling set (80%–20%), which is one of the most common and accurate procedures used to evaluate these types of algorithms [25, 40].

For the fivefold cross-validation and external test set, values between 78.5 and 87.7% were obtained. Due to the characteristics of the dataset, for all models, the true-negative rates, indicating the identification of inactive molecules (ranging from 78.5 to 91.0%), were greater than the true-positive rates, indicating the identification of active molecules (from 75.7 to 87.7%). These outcomes demonstrated that the generated classification models were highly restrictive, minimizing the probability of obtaining false-positive structures and preventing inactive molecules from being predicted to be active. Additionally, the models that were built with the Decision Tree Learner algorithm returned lower positive, negative, and overall rates (Table 1). Two additional models were built using

Table 1 Summary of the fivefold cross-validation, which was obtained using three algorithms and three types of molecular descriptors on a total set of 7032 compounds, with inhibitory *in vitro* activity against AChE

	Fivefold Cross-validation				External test set		
	Algorithm	Active	Inactive	Overall	Active	Inactive	Overall
Volsurf+	*DTL*	76.2	80.4	78.5	75.7	78.5	77.3
	GBT	79.4	84.1	82.0	77.4	85.1	81.7
Bit-based Fingerprint	*DTL*	85.2	86.9	86.2	87.7	84.8	86.1
	GBT	79.4	90.5	85.6	79.4	91.0	85.8
	SVM	84.6	86.6	85.7	86.4	86.7	86.6
DRAGON 5.0	*DTL*	81.5	84.4	83.1	83.4	86.2	84.9
	GBT	85.1	89.8	87.7	84.1	90.2	87.4

DTL = Decision Tree Learner, *GBT* = Gradient Boosted Tree Learner, and *SVM* = Support Vector Machine

a combination of VolSurf+ and DRAGON 5.0 molecular descriptors using the Decision Tree Learner and Gradient Boosted Tree Learner algorithms, which returned similar rates as those models that were built using molecular descriptors in isolation (Supplementary Material).

From the confusion matrix, two quality parameters were calculated: AUC and MCC. Because an imbalanced dataset was generated using chosen cutoff value ($pIC_{50} \geq 6.0$), the SMOTE technique was applied to redistribute the imbalanced dataset. SMOTE is a technique for generating new synthetic instances of an outcome to re-balance class distributions. SMOTE creates extra synthetic instances by pairing each positive instance with its nearest neighbor, then generating new instances s along a line segment between the pair of positive instances [33].

The ROC curve is a quality parameter that plots the true-positive rate (sensitivity, Eq. 1) against the false-positive rate (1− specificity, Eq. 2), and the values for the AUC can range between 0 and 1, with a value of 1 indicating a perfect separation between the two groups) [34].

$$\text{Sensitivity} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{1}$$

$$\text{Specificity} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{2}$$

For all classification models obtained, AUC values greater than 0.80 were achieved, demonstrating a high rate of sensitivity and a low false-positive rate, and the classification models built with the Gradient Boosted Tree Learner algorithm resulted in a higher degree of differentiation compared with the other types of models, presenting AUC values above of 0.90 (Fig. 2).

The other quality parameter calculated was the MCC, which was determined from all of the values obtained from the confusion matrix, as shown in Eq.

$$\text{MCC} = \frac{(\text{TP} \cdot \text{TN}) - (\text{FP} \cdot \text{FN})}{\sqrt{(\text{TP} + \text{FP})(\text{TP} + \text{FN})(\text{TN} + \text{FP})(\text{TN} + \text{FN})}} \tag{3}$$

where TP is the true-positive rate, TN is the true-negative rate, FP is the false-positive rate, and FN is the false-negative rate. An MCC value equal to 1 indicates a perfect correlation, a value of 0 indicates a random prediction, and a value of −1 indicates total disagreement between the prediction and the observation [35].

MCC values from 0.57 to 0.75 and between 0.54 and 0.75 were observed for the fivefold cross-validation and external test sets, respectively. Additionally, in those models built using DRAGON 5.0 and VolSurf+ molecular descriptors, the MCC values were lower for models built using the Decision Tree Learner algorithm compared with the values for the models built with the Gradient Boosted Tree Learner algorithm, which indicated that the Gradient Boosted Tree Learner models had a lower false-positive rate (below 15.9%), as indicated by the AUC results, but also had a higher hit rate for inactive compounds. For bit-based fingerprint descriptors, the models built with the Decision Tree Learner algorithm exhibited similar MCC values for the fivefold cross-validation and external test sets (0.72 and 0.72, respectively) compared with the models built using the Gradient Boosted Tree Learner (0.71 and 0.71, respectively) and Support Vector Machine (0.71 and 0.73, respectively) algorithms. Similarly, higher MCC values were observed for the models built with bit-based fingerprints, which were associated with MCC values above 0.70, except for the model that used the Gradient Boosted Tree Learner algorithm and DRAGON 5.0 molecular descriptors, which showed MCC values of 0.75 for both the fivefold cross-validation and external test sets. All analyses were performed after the redistribution of the imbalanced dataset using SMOTE.

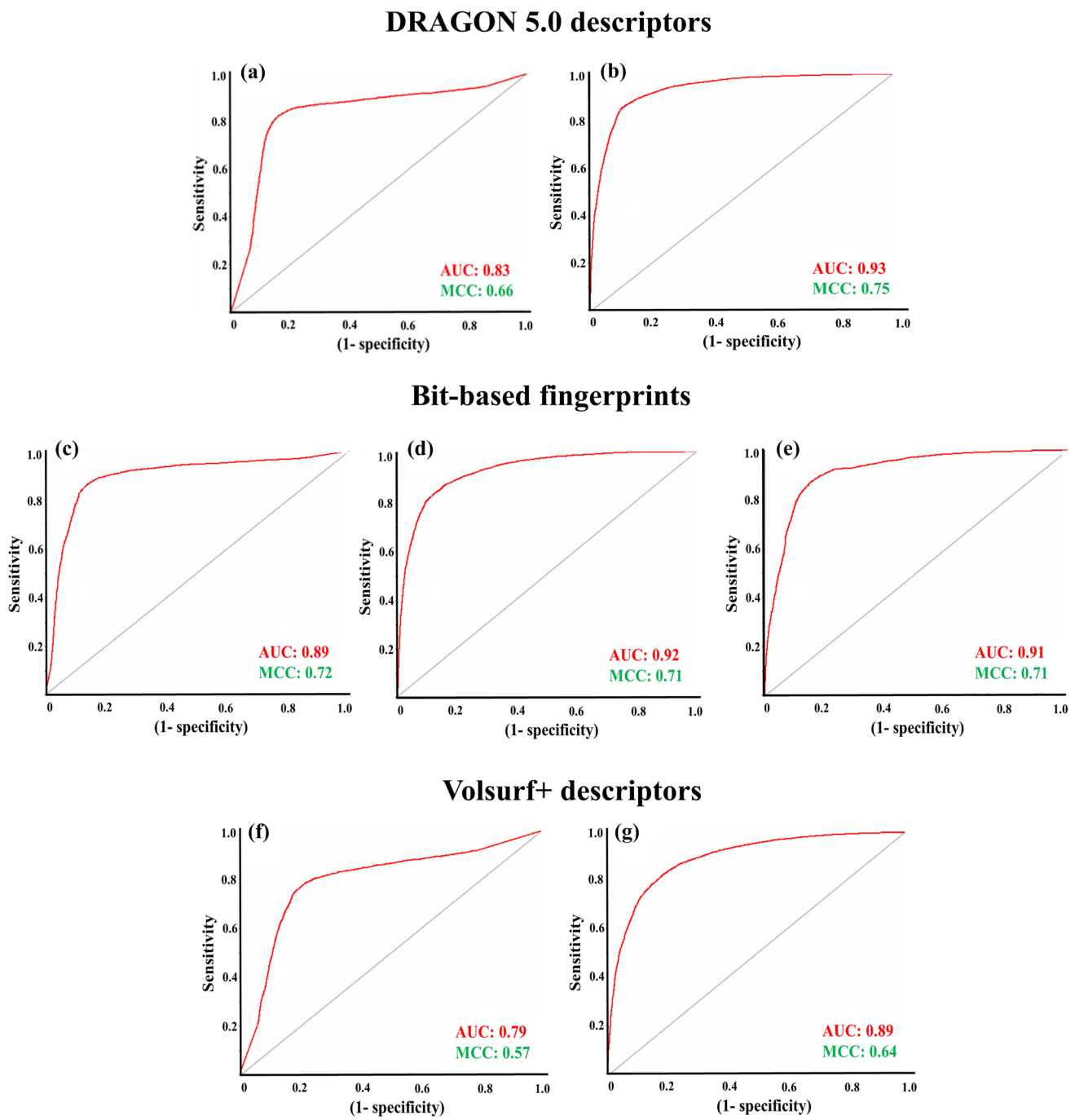


Fig 2 ROC plots, sensitivity versus [1—specificity], generated for machine learning classification models (fivefold cross-validation) that were built using three different types of molecular descriptors. The ROC curves were built after the data set imbalance was redistributed using SMOTE. **a, c, and d** DTL, Decision Tree Learner algorithm; **b, d**, and **f** GBT, Gradient Boosted Tree Learner algorithm; and **g** SVM, Support Vector Machine algorithm; AUC = value of the area under the curve; MCC = Matthews's correlation coefficient

To identify any compounds in the test sets and the SistematX dataset for which the predictions may be unreliable, the APD was calculated. In the seven generated classification models, less than 0.7% of the test set was classified as unreliable. Similarly, for the SistematX dataset used for the ligand-based virtual screening, more than 97.7% of the secondary metabolites were classified as reliable.

Ligand-based virtual screening of SistematX database

All of the generated machine learning classification models were used to analyze the 8,953 secondary metabolites registered in SistematX to predict those that may potentially have inhibitory activity against human AChE, the inhibition of which is considered to be a promising strategy for the treatment of neurological disorders, such as Alzheimer's disease, senile dementia, ataxia, and myasthenia gravis [41].

The probability of each secondary metabolite to be classified as active, ***p(A),*** was calculated through a consensus analysis of all probability values obtained from the fivefold classification models, according to Eq. 4:

$$\boldsymbol{p(A)} = \frac{\sum_{i=1} \left(p_i \cdot \mathrm{TN_i}\right)}{\sum_{i=1} \mathrm{TN_i}} \tag{4}$$

where ***p(A)*** is the combined probability of all machine learning classification models; $\boldsymbol{p_i}$ is the probability value calculated in the machine learning classification model *(i)*; TN is the true-negative rate (Specificity, Eq. 2) in the machine learning classification model *(i)*, and i ranges from 1 to 8593 (SistematX dataset).

A consensus analysis was selected because one of the main objectives of this study was to select various types of molecular descriptors that reflect the different aspects of the molecular structure to minimize the false-positive rate. According to Gramatica et al., selecting the best performing models may overemphasize some aspects of the molecular structure and underestimate others or result in the complete bypassing of many important features. The consensus analysis provided better prediction outcomes than the majority of the individual models and might consider more peculiar aspects of certain structures [42].

After this calculation, only 25 of the 8,593 secondary metabolites registered in the SistematX webtool achieved ***p(A)*** values equal to or greater than 0.5. Four types of secondary metabolites were observed among the structures classified as active (16 flavonoids, 4 sesquiterpene lactones, 4 diterpenes, and 1 monoterpene), which have been identified in species from three botanical families: Asteraceae, Annonaceae, and Apocynaceae (Fig 3 and Table 2).

Structural features were identified, such as the presence of glycosides containing a disaccharide at the C3 position of the hydroxyl group on carbon-3 in the flavonols **1** and **2**, which are two secondary metabolites that were extracted from Asteraceae and are highly distributed throughout the American continent, such as *Carthamus* and *Brickellia* [43, 44]. However, after applying the PAINS remover web tool to examine the results for false-positive molecules, 9 of the 25 structures classified as active against AChE were classified as false positives due to the presence of the catechol moiety in their structures [45].

Similarly, the moiety α-methylene-γ-lactone was observed in three of the four sesquiterpene lactones classified as active compounds (structures **6**, **19,** and **24**). Previously, the presence of this group has been associated with antiparasitic activity in this type of metabolite, due to the interaction between this group and the sulfhydryl group of cysteine, through a Michael addition [46]. Similarly, a heliangolide skeleton was present in two sesquiterpene lactones (structures **6** and **24**). These two secondary metabolites were identified in species belonging to the subtribe Liatrinae of the Asteraceae family [47].

Molecular docking calculations

In addition, molecular docking calculations were performed to establish a consensus, with results obtained from machine learning classification models. The crystal structure of recombinant human AChE (PDB ID: 4EY6) in complex with (–)-galantamine (PDB ID: GLN), which is a reversible, competitive, tertiary alkaloid AChE inhibitor, was obtained from the PDB databank [3, 36]. The molecular docking protocol was designed to identify exclusive secondary metabolites that utilize the same inhibition mechanism against AChE as galantamine. The methodology was validated by performing redocking with the ligand reported in the PDB crystal structure.

Using the same parameters, virtual screening of the 8,593 structures found in the SistematX dataset was performed. Based on the binding energy values, all tested molecules were ranked using the following probability calculation (Eq. 5) [48]:

$$\boldsymbol{p(D)} = \frac{E_i}{E_{\mathrm{min}}} \quad \text{if } E_i < E_{\mathrm{galantamine}} \tag{5}$$

where ***p(D)*** = molecular docking probability; E_{i} = docking energy of compound i, for which i ranges from 1 to 8593 (SistematX dataset); E_{min} = the lowest energy value of the dataset; $E_{\mathrm{galantamine}}$ = the galantamine energy from the protein crystallography redocking attempt.

The number of secondary metabolites with ***p(D)*** values greater than 0.5 and binding energy values less than the binding energy of galantamine (−129.5 kJ/mol) was 2244 (26.1%). The root-mean-square deviation (RMSD) for galantamine was 0.0656. Of the 16 secondary metabolites from the SistematX dataset identified as potential AChE inhibitors by the machine learning models, only 8 structures presented ***p(D)*** values greater than 0.5, and structure **11,** a flavonol derived from Asteraceae, was the top-ranked structure, with a ***p(D)*** value of 0.77 and a docking score of −172.9 kJ/mol.

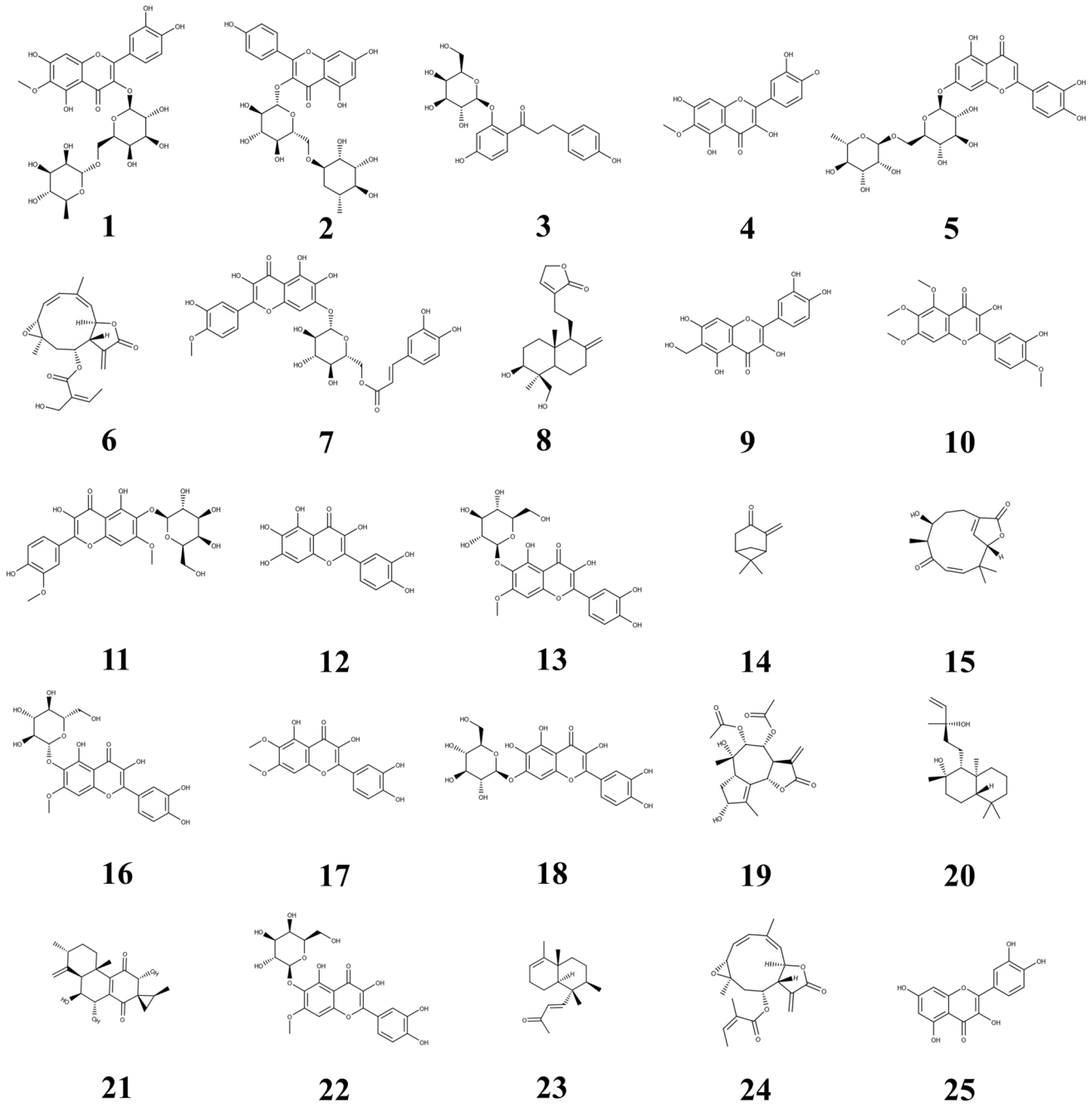


Fig 3. Secondary metabolites from SistematX were classified as potential inhibitors of AChE by the developed machine learning classification models

Analyzing the docking poses, galantamine showed 20 interactions with amino acids within the AChE active site. Only one hydrogen bond (H-bond) was formed between the oxygen present in the aromatic ring of galantamine and S203. The interactions between the structures **3, 11, 15,** and **24** and AChE were analyzed and compared against the interaction observed between galantamine and AChE. More than 65% of the galantamine-interacting residues identified in AChE (15) also interacted with four secondary metabolites; however, the interaction with the residue W236 was only observed for galantamine. All interactions between AChE and galantamine and the four selected secondary metabolites are shown in Fig. 4.

The selected secondary metabolites showed a higher number of H-bonds compared with those observed for galantamine, and structure **3** (chalcone) presented the highest number of H-bond interactions with W86, G120, G121, G122, and S203. Additionally, for both evaluated flavonoids,

Table 2 Botanical data and ***p(A)*** values for the ten molecules classified as active molecules, identified by virtual screening. Secondary metabolites highlighted in bold were classified as false-positive structures by the PAINS remover [45]

ID	Class	Skeleton	Family	Species	p(A)
1	**Flavonoid**	**Flavonol**	**Asteraceae**	***Brickellia arguta***	**0.61**
2	Flavonoid	Flavonol	Asteraceae	*Carthamus tinctorius*	0.60
3	Flavonoid	Chalcone	Asteraceae	*Blumea balsamifera*	0.59
4	**Flavonoid**	**Flavonol**	**Asteraceae**	***Achillea biebersteinii***	**0.59**
5	**Flavonoid**	**Flavone**	**Asteraceae**	***Conyza bonariensis***	**0.58**
6	Sesquiterpene lactone	Heliangolide	Asteraceae	*Hartwrightia floridana*	0.57
7	Flavonoid	Flavonol	Asteraceae	*Eupatorium glandulosum*	0.57
8	Diterpene	Andrographolide-type	Lamiaceae	*Andrographis paniculata*	0.57
9	**Flavonoid**	**Flavonol**	**Asteraceae**	***Tagetes dianthioflora***	**0.57**
10	Flavonoid	Flavonol	Asteraceae	*Eupatorium semiserratum*	0.56
11	Flavonoid	Flavonol	Asteraceae	*Tagetes mandonii*	0.56
12	Flavonoid	Flavonol	Asteraceae	*Tagetes dianthioflora*	0.56
13	**Flavonoid**	**Flavonol**	**Asteraceae**	***Tagetes minuta***	**0.56**
14	Monoterpene	Pinane	Annonaceae	*Annona reticulata*	0.56
15	Sesquiterpene lactone	Humulene	Asteraceae	*Asteriscus graveolens*	0.55
16	**Flavonoid**	**Flavonol**	**Asteraceae**	***Tagetes minuta***	**0.55**
17	**Flavonoid**	**Flavonol**	**Asteraceae**	***Brickellia dentata***	**0.55**
18	Flavonoid	Flavonol	Asteraceae	*Eupatorium adenophorum*	0.54
19	Sesquiterpene lactone	Guaianolide	Asteraceae	*Anthemis cretica*	0.54
20	Diterpene	–	Lamiaceae	*Vitex negundo*	0.54
21	Diterpene	–	Lamiaceae	*Coleus somaliensis*	0.54
22	**Flavonoid**	**Flavonol**	**Asteraceae**	***Tagetes mandonii***	**0.54**
23	Bisnorditerpene	-	Annonaceae	*Polyalthia viridis*	0.54
24	Sesquiterpene lactone	Heliangolide	Asteraceae	*Liatris punctata*	0.53
25	**Flavonoid**	**Flavonol**	**Asteraceae**	***Tagetes dianthioflora***	**0.52**

structures **3** (chalcone) and **8** (flavonol), two π–π interactions were observed between the aromatic rings of these molecules and the tyrosine residues (Y124 and Y337) of AChE. In structure **11,** a higher number of carbon–H-bond interactions was observed, with fewer carbon–H-bond interactions observed for structures **3** and **15,** and no instances of this interaction type observed in structure **6** or with galantamine.

The residue S203 also interacts with the selected structures, and an H-bond was formed with structures **3** (chalcone), **15,** and **24** (two sesquiterpene lactones). Structure **11** established a van der Waals interaction with S203. Similarly, all structures interacted with G122 through an H-bond for structures **3**, **11,** and **15** (Fig. 4), whereas galantamine and structure **24** interacted with this residue through van der Waals interactions. Both sesquiterpene lactones (structures **15** and **24**) interacted through the carbonyl group present in the lactone ring, establishing H-bond interactions with G122 and Y337, respectively.

Flexible molecular docking calculations were also performed for the best-ranked molecules to analyze the behaviors of these molecules at the active site of the target in more detail by mimicking the natural biological environment. Docking results should be viewed in a theoretical context because they are not supported by experimental evidence [49]. Table 3 shows the most important residues involved in the interactions with the studied target, emphasizing the interactions with those residues that appeared to be critical for binding with AChE in the structure-based virtual screening analysis.

The eight selected secondary metabolites established favorable interactions with residues W86, G122, and E202, and W86 showed the strongest binding energy values (from −36.2 to −5.0 kJ/mol). The flavonols **11** and **12** exhibited similar docking values for the evaluated residues, differing only in the contributions to interactions with residues D74 and Y341. For the sesquiterpene lactones, different behaviors were identified. Structure **24** interacts more favorably with key residues in the pocket of AChE, whereas guaianolide (structure **15**) was not observed to interact with D74 and Y341 and featured an unfavorable interaction with S203 (17.6 kJ/mol).

Consensus analysis of machine learning classification models and molecular docking calculations

Consensus analysis of the two methodologies used in this study (machine learning and molecular docking) was

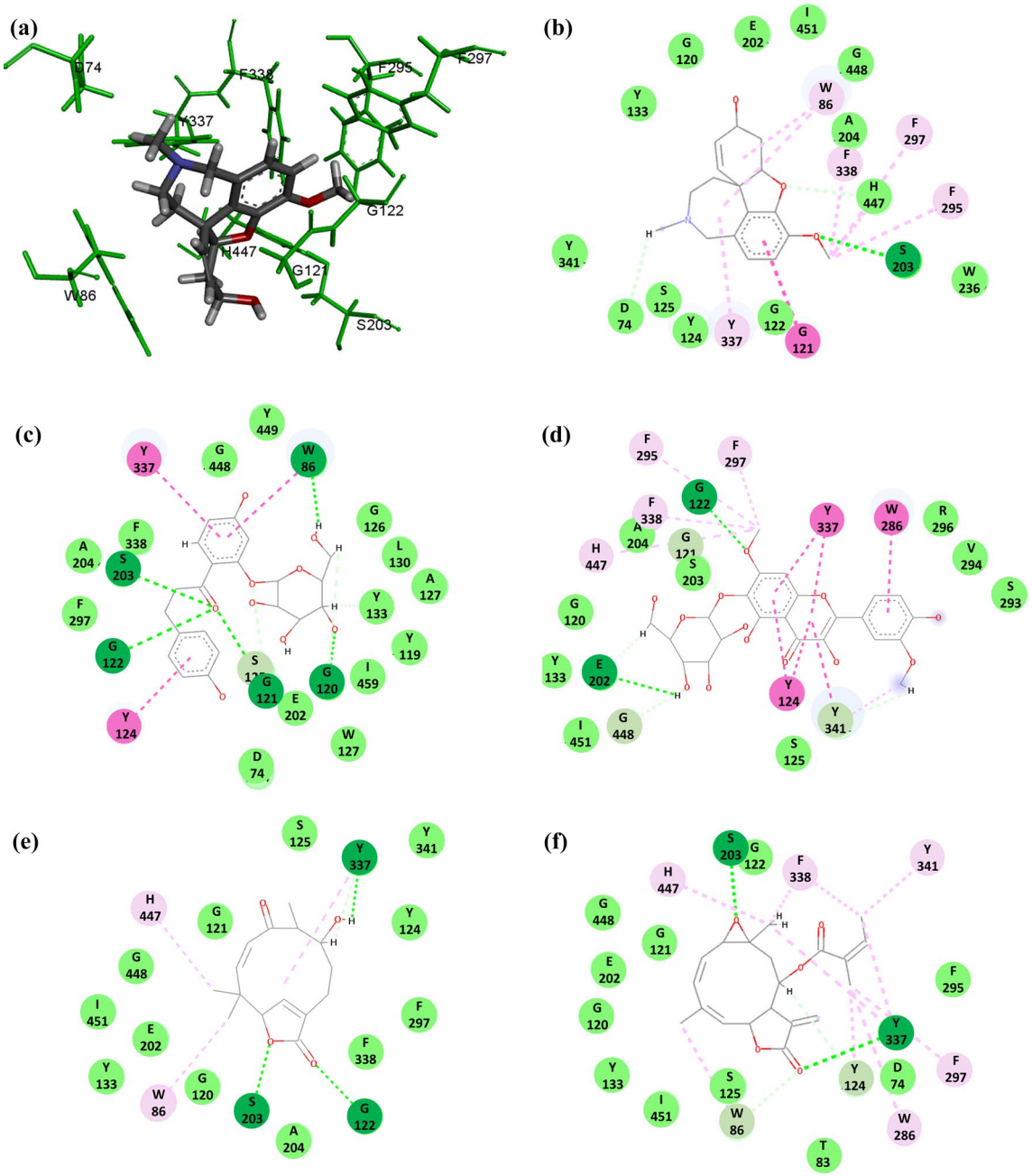


Fig 4 **a** Docking conformation of galantamine in the active site of human AChE (Green); 2D-residual interaction diagrams of **b** galantamine, **c** Structure 3, **d** Structure 11, **e** Structure 15, and **f** Structure 24. The interacting residues are shown as colored circles, and interactions are indicated as colored dashed lines: H-bond (lime), van der Waals (green), π–π (purple) and π–alkyl (pink), unfavorable (red) and carbon–H-bond (teal)

performed to verify potentially active secondary metabolites with AChE inhibitory activity and explore their inhibitory mechanisms. A new probability score, ***p(AChE)***, was determined, combining the probability scores of ***p(A)*** and ***p(D)*** (Eq. 6). An additional validation procedure for molecular docking calculations was performed using the entire AChE dataset (7,032 structures) in an attempt to minimize the probability of selecting false-positive compounds..

$$p(AChE) = \frac{p(D) \cdot \left(\mathrm{TN}_{\mathrm{docking}}\right) + p(A)}{1 + \mathrm{TN}_{\mathrm{docking}}} \tag{6}$$

where ***p(AChE)*** = combined probability; ***p(D)*** = molecular docking probability; ***p(A)*** = probability obtained from

Table 3 Interaction values (kJ/mol) for eight selected secondary metabolites identified in the SistematX dataset (classified as active in the machine learning classification models and by the molecular docking calculations) and selected AChE residues

Structure	D74	E202	G120	G121	G122	S203	W86	Y124	Y337	Y341
2	−8.5	−9.4	−1.5	−4.1	−3.1	19.5	−5.0	42.9	−23.8	−21.9
3	−7.3	−7.5	−6.2	−3.2	−6.1	29.4	−30.8	14.8	13.2	−5.2
6	−6.0	−5.4	3.4	4.8	−2.1	10.5	−15.6	14.6	9.9	−8.2
11	−4.7	−7.1	3.5	−7.1	−2.4	−2.8	−19.4	−5.5	−21.3	9.8
12	9.2	−4.8	−1.6	−6.3	−0.9	−4.1	−19.4	−5.0	−21.3	−1.4
15	–	−7.6	−3.9	−12.4	−6.9	17.6	−11.7	−6.0	−7.6	–
19	−3.6	−8.1	9.7	0.8	−3.5	2.7	−36.2	5.0	18.7	−5.8
24	−4.7	−7.1	3.5	−7.1	−2.4	−2.8	−19.4	−5.5	−21.3	9.8
GLN	−1.9	−8.0	−4.6	−14.9	−6.4	−7.9	−25.4	−6.9	−12.1	−1.4

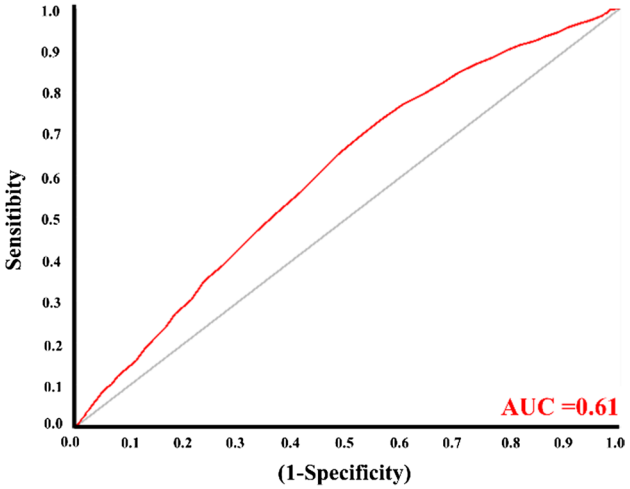


Fig 5 Validation of the molecular docking calculations using an ROC curve depicting sensitivity versus (1—specificity). AUC: area under the curve

machine learning classification models; $TN_{Docking}$ = the true-negative rate (specificity, Eq. 2) obtained from the validation of the molecular docking calculations.

To calculate the true-negative rate from the docking calculations, the docking scores for all structures were calculated, and their ***p(D)*** values were determined according to Eq. 5 (molecular docking probabilities). The structures were then classified as either active or inactive to obtain two proportional structure groups. The cutoff selected for this procedure was the docking score for galantamine (−129.5 kJ/mol) + 15 kJ/mol. As a result, 3454 (49.1%) molecules were classified as active, and 3578 (50.9%) molecules were classified as inactive. Using the ***p(D)*** values and the true-positive rate, the confusion matrix and its respective ROC curve were determined (Fig. 5.).

The true-negative rate (specificity, Eq. 2), which was obtained from the validation of the molecular docking calculation, is related to Eq. 6, similar to the relationship observed with Eq. 4 (ligand-based probabilities) for the fivefold cross-validation machine learning models. The goal of this analysis was to minimize the probability of selecting inactive molecules as active molecules (false-positive) because the selection of false-positive molecules can result in significant wastes of both time and money [23, 48].

Table 4 Secondary metabolites identified as potential inhibitors of AChE through a consensus analysis of machine learning classification models and molecular docking calculations

Structure	*p(A)*	*p(D)*	*p(AChE)*
11	0.56	0.77	0.64
2	0.60	0.69	0.63
3	0.59	0.70	0.63
12	0.56	0.70	0.61
6	0.57	0.65	0.60
19	0.54	0.67	0.59
24	0.53	0.61	0.56
15	0.55	0.58	0.56

Table 4 summarizes the results for the best-ranked secondary metabolites according to the consensus analysis. Only eight structures were identified as potential AChE inhibitors from among the 8,593 secondary metabolites registered in SistematX, which was used as the starting point for this study. Four sesquiterpene lactones and four flavonoids from Asteraceae were the unique molecules that achieved ***p(A)***, ***p(D),*** and ***p(AChE)*** values above 0.5.

Drug-like properties of the potential inhibitors against AChE

To be effective as therapeutic agents against AChE, centrally acting drugs must be able to cross the BBB. SwissADME [37], was used to evaluate the qualitative capacity of each potential inhibitor to cross the BBBs. The obtained results were evaluated in a quantitative model using VolSurf+, containing approximately 500 related but chemically diverse compounds extracted from the literature and our in-house data set that are known to be brain-penetrating [27, 28, 39]

Table 5 Drug-like properties of secondary metabolites identified as potential AChE inhibitors

ID	Molecular weight[a]	BBB permeant[a]	LgBB[b]	GI absorption[a]	Mutagenic[c]	Tumorigenic[c]	Reproductive Effective [c]	Irritant[c]	Lipinski violation[a]	Veber violation [a]
2	592.55	No	−4.83	Low	None	None	None	None	3	1
3	420.41	No	−2.14	Low	None	None	None	None	1	1
6	360.4	No	−0.41	High	None	None	None	High	0	0
11	508.43	No	−2.86	Low	None	None	None	None	3	1
12	318.24	No	−3.23	Low	None	None	None	None	1	1
15	264.32	Yes	−0.22	High	None	None	None	None	0	0
19	380.39	No	−1.24	High	None	None	None	High	0	0
24	344.4	Yes	0.02	High	None	None	None	High	0	0

p(D) = molecular docking probability; *p(A)* = probability obtained from machine learning classification models; *p(AChE)* = combined probability value

[a]SwissADME

[b]Volsurf+

[c]Osiris Data Warrior v.5.2.1

In the qualitative prediction, structures **15** and **24** were identified as having the potential to cross the BBB, indicating the potential for neuroprotective effects. This result was confirmed for structure **24** in the VolSurf + model, with an LgBB score of 0.02, which is classified as a compound with moderate BBB permeation. Structure **15** (−0.22) possessed a minimal ability to cross the BBB (Table 5).

Lipinski's "rule of five" and the Veber rules were evaluated for the four sesquiterpene lactones obtained from Asteraceae (structures **6**, **15**, **19,** and **24**), and no violations were identified, which suggested that these secondary metabolites are likely orally bioavailable and demonstrated high gastrointestinal absorption. Moreover, both of the identified flavonol glycosides (structures **2** and **11**) showed multiple violations of the Lipinski and Veber rules [37].

Finally, mutagenicity, tumorigenesis, negative effects on the reproductive system, and irritability were evaluated for these four compounds also were determined using OSIRIS Data Warrior v.5.2.1[38] . Only the sesquiterpenes lactones (structures **6**, **19,** and **24**) showed negative effects, which were predicted to cause high irritability (Table 5). Overall, the results showed that structures **15** and **24** were highly likely to be easily absorbed and distributed, with low levels of toxicity. However, irritability was predicted for structure **24**.

Conclusions

This study combined machine learning classification models with molecular docking calculations to assess 8593 secondary metabolites from the SistematX database as potential AChE inhibitors. For the construction of the classificatory models, we utilized a dataset of 7032 structures registered in BindingDB and ChEMBL that have previously reported *in vitro* inhibitory activities against AChE. The models obtained achieved an accuracy above 78.5% and were highly restrictive. Only 25 molecules were classified as being potentially active. Some shared structural features were observed among these molecules (nine flavonoids were identified as false positives by the PAINS server). Molecular docking allows the predicted interactions to be observed between the secondary metabolites and the active site of AChE.

Combining the probability values obtained for each stage, a consensus analysis was performed, and eight structures (four flavonoids and four sesquiterpene lactone) were identified as potential inhibitors of AChE that utilize the same mechanism of action as galantamine. Finally, through the prediction of ADME, toxicity, and pharmacokinetic properties of the four potentially active molecules against diverse diseases, structures **15** and **24** emerged as the most promissory secondary metabolites against AChE among all 8593 molecules tested.

These methodologies that integrate multiple virtual screening approaches represent interesting alternatives for use as the starting point during drug development, allowing for the identification of potentially actives molecules against diverse diseases. In addition, the use of other computational tools, such as molecular docking calculations and ADMET predictions, allows for the establishment of a predicted mechanism of action for the selected structures, reducing costs and saving time and money.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s11030-021-10245-z.

Funding We thank the CNPq and Capes for financial Support, Grant Numbers 309648/2019-0 and 431254/2018-4.

Declarations

Conflict of interest The authors declare that they have no conflict of interest.

References

1. Taylor P, Camp S, Radić Z (2009) Acetylcholinesterase. In: Squire LR (ed) Encyclopedia of neuroscience. Academic Press, Oxford, pp 5–7
2. Shannon MW, Borron SW, Burn MJ (2007) Chemical weapons. Haddad and Winchester's clinical management of poisoning and drug overdose (Fourth Edition). W.B. Saunders, Philadelphia, pp 1487–520
3. Cheung J, Rudolph MJ, Burshteyn F, Cassidy MS, Gary EN, Love J et al (2012) Structures of human acetylcholinesterase in complex with pharmacologically important ligands. J Med Chem 55(22):10282–6. https://doi.org/10.1021/jm300871x
4. Rotundo RL (2003) Expression and localization of acetylcholinesterase at the neuromuscular junction. J Neurocytol 32(5):743–66. https://doi.org/10.1023/B:NEUR.0000020621.58197.d4
5. Acetylcholinesterase BD (2007). In: Enna SJ, Bylund DB (eds) xPharm: the comprehensive pharmacology reference. Elsevier, New York, pp 1–8
6. Akıncıoğlu H, Gülçin İ (2020) Potent acetylcholinesterase inhibitors: potential drugs for Alzheimer's disease. Mini Rev Med Chem 20(8):703–15. https://doi.org/10.2174/1389557520666200103100521
7. Mirjana BC, Danijela ZK, Tamara DL-P, Aleksandra MB, Vesna MV (2013) Acetylcholinesterase inhibitors: pharmacology and toxicology. Curr Neuropharmacol 11(3):315–335. https://doi.org/10.2174/1570159X11311030006
8. Poddar MK, Banerjee S, Chakraborty A, Dutta D (2021) Metabolic disorder in Alzheimer's disease. Metab Brain Dis. https://doi.org/10.1007/s11011-021-00673-z
9. Devita M, Masina F, Mapelli D, Anselmi P, Sergi G, Coin A (2021) Acetylcholinesterase inhibitors and cognitive stimulation, combined and alone, in treating individuals with mild Alzheimer's disease. Aging Clin Exp Res. https://doi.org/10.1007/s40520-021-01837-8
10. Chekmarev D, Kholodovych V, Kortagere S, Welsh W, Ekins S (2009) Predicting inhibitors of acetylcholinesterase by regression and classification machine learning approaches with combinations of molecular descriptors. Pharm Res 26:2216–24. https://doi.org/10.1007/s11095-009-9937-8
11. Leuci R, Brunetti L, Poliseno V, Laghezza A, Loiodice F, Tortorella P et al (2021) Natural compounds for the prevention and treatment of cardiovascular and neurodegenerative diseases. Foods 10(1):29. https://doi.org/10.3390/foods10010029
12. Newman DJ, Cragg GM (2020) Natural products as sources of new drugs over the nearly four decades from 01/1981 to 09/2019. J Nat Prod 83(3):770–803. https://doi.org/10.1021/acs.jnatprod.9b01285
13. Calixto JB (2019) The role of natural products in modern drug discovery. Anais da Academia Brasileira de Ciências. https://doi.org/10.1590/0001-3765201920190105
14. Afendi FM, Okada T, Yamazaki M, Hirai-Morita A, Nakamura Y, Nakamura K et al (2012) KNApSAcK family databases: integrated metabolite–plant species databases for multifaceted plant research. Plant Cell Physiol 53(2):e1. https://doi.org/10.1093/pcp/pcr165
15. Banerjee P, Erehman J, Gohlke B-O, Wilhelm T, Preissner R, Dunkel M (2015) Super Natural II—a database of natural products. Nucl Acids Res 43(D1):D935–D9. https://doi.org/10.1093/nar/gku886
16. Sorokina M, Merseburger P, Rajan K, Yirik MA, Steinbeck C (2021) COCONUT online: Collection of Open Natural Products database. J Cheminformat 13(1):1–13. https://doi.org/10.1186/s13321-020-00478-9
17. Valli M, Dos Santos RN, Figueira LD, Nakajima CH, Castro-Gamboa I, Andricopulo AD et al (2013) Development of a natural products database from the biodiversity of Brazil. J Nat Prod 76(3):439–44. https://doi.org/10.1021/np3006875
18. Scotti MT, Herrera-Acevedo C, Oliveira TB, Costa RPO, Santos SYKdO, Rodrigues RP et al (2018) SistematX, an online web-based cheminformatics tool for data management of secondary metabolites. Molecules 23(1):103. https://doi.org/10.3390/molecules23010103
19. Thomford NE, Senthebane DA, Rowe A, Munro D, Seele P, Maroyi A et al (2018) Natural products for drug discovery in the 21st century: innovations for novel drug discovery. Int J Mol Sci 19(6):1578. https://doi.org/10.3390/ijms19061578
20. Liu T, Lin Y, Wen X, Jorissen RN, Gilson MK (2007) BindingDB: a web-accessible database of experimentally determined protein–ligand binding affinities. Nucl Acids Res 35(suppl_1):D198–D201. https://doi.org/10.1093/nar/gkl999
21. Atanasova M, Yordanov N, Dimitrov I, Berkov S, Doytchinova I (2015) Molecular docking study on galantamine derivatives as cholinesterase inhibitors. Mol Informat 34(6–7):394–403. https://doi.org/10.1002/minf.201400145
22. Stavrakov G, Philipova I, Zheleva-Dimitrova D, Valkova I, Salamanova E, Konstantinov S et al (2017) Docking-based design and synthesis of galantamine–camphane hybrids as inhibitors of acetylcholinesterase. Chem Biol Drug Design 90(5):709–18. https://doi.org/10.1111/cbdd.12991
23. Herrera-Acevedo C, Maia MDS, Cavalcanti ÉBVS, Coy-Barrera E, Scotti L, Scotti MT (2011) Selection of antileishmanial sesquiterpene lactones from SistematX database using a combined ligandstructure-based virtual screening approach. Mol Divers. https://doi.org/10.1007/s11030-020-10139-6
24. Cherkasov A, Muratov EN, Fourches D, Varnek A, Baskin II, Cronin M et al (2014) QSAR modeling: where have you been? Where are you going to? J Med Chem 57(12):4977–5010. https://doi.org/10.1021/jm4004285
25. Fourches D, Muratov E, Tropsha A (2015) Curation of chemogenomics data. Nature Chem Biol 11(8):535. https://doi.org/10.1038/nchembio.1881
26. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. J Mol Biol 215(3):403–10. https://doi.org/10.1016/S0022-2836(05)80360-2
27. Cruciani G, Crivori P, Carrupt PA, Testa B (2000) Molecular fields in quantitative structure–permeation relationships: the VolSurf approach. J Mol Struct: THEOCHEM 503(1–2):17–30. https://doi.org/10.1016/S0166-1280(99)00360-7
28. Cruciani G, Pastor M, Guba W (2000) VolSurf: a new tool for the pharmacokinetic optimization of lead compounds. Eur J Pharm Sci 11:S29–S39. https://doi.org/10.1016/S0928-0987(00)00162-7
29. Mauri A, Consonni V, Pavan M, Todeschini R (2006) Dragon software: an easy approach to molecular descriptor calculations. Match 56(2):237–48
30. Todeschini R, Consonni V (2008) Handbook of molecular descriptors. Wiley, New York
31. Scotti L, Fernandes MB, Muramatsu E, Emereciano VdP, Tavares JF, Silva MSd et al (2011) 13C NMR spectral data and molecular descriptors to predict the antioxidant activity of flavonoids. Braz

J Pharm Sci 47(2):241–249. https://doi.org/10.1590/S1984-82502011000200005
32. Berthold MR, Cebron N, Dill F, Gabriel TR, Kötter T, Meinl T et al (2009) KNIME-the Konstanz information miner: version 2.0 and beyond. AcM SIGKDD Explor Newslett 11(1):26–31. https://doi.org/10.1145/1656274.1656280
33. Siriseriwan W, Sinapiromsaran K (2016) The effective redistribution for imbalance dataset: relocating safe-level SMOTE with minority outcast handling. Chiang Mai J Sci 43(1):234–46
34. Hanley JA, McNeil BJ (1982) The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology 143(1):29–36. https://doi.org/10.1148/radiology.143.1.7063747
35. Matthews BW (1975) Comparison of the predicted and observed secondary structure of T4 phage lysozyme. Biochimica et Biophysica Acta (BBA)-Protein Struct 405(2):442–451. https://doi.org/10.1016/0005-2795(75)90109-9
36. Scott LJ, Goa KL (2000) Galantamine. Drugs 60(5):1095–1122
37. Daina A, Michielin O, Zoete V (2017) SwissADME: a free web tool to evaluate pharmacokinetics, drug-likeness and medicinal chemistry friendliness of small molecules. Sci Rep 7(1):1–13. https://doi.org/10.1038/srep42717
38. Sander T, Freyss J, von Korff M, Rufener C (2015) DataWarrior: an open-source program for chemistry aware data visualization and analysis. J Chem Inf Model 55(2):460–73. https://doi.org/10.1021/ci500588j
39. Crivori P, Cruciani G, Carrupt P-A, Testa B (2000) Predicting blood—brain barrier permeation from three-dimensional molecular structure. J Med Chem 43(11):2204–16. https://doi.org/10.1021/jm990968+
40. Wong T-T, Yeh P-Y (2019) Reliable accuracy estimates from k-fold cross validation. IEEE Trans Knowl Data Eng 32(8):1586–94. https://doi.org/10.1109/TKDE.2019.2912815
41. Mukherjee PK, Kumar V, Mal M, Houghton PJ (2007) Acetylcholinesterase inhibitors from plants. Phytomedicine. 14(4):289–300. https://doi.org/10.1016/j.phymed.2007.02.002
42. Gramatica P, Giani E, Papa E (2007) Statistical external validation and consensus modeling: a QSPR case study for Koc prediction. J Mol Gr Model 25(6):755–66. https://doi.org/10.1016/j.jmgm.2006.06.005
43. Mues R, Timmermann BN, Ohno N, Mabry TJ (1979) 6-Methoxyflavonoids from *Brickellia californica*. Phytochemistry 18(8):1379–83. https://doi.org/10.1016/0031-9422(79)83027-7
44. Bohm BA, Stuessy TF (2001) Flavonoids of the sunflower family (Asteraceae). Springer, Vienna
45. Baell JB, Holloway GA (2010) New substructure filters for removal of pan assay interference compounds (PAINS) from screening libraries and for their exclusion in bioassays. J Med Chem 53(7):2719–40. https://doi.org/10.1021/jm901137j
46. Schmidt TJ (1999) Toxic activities of sesquiterpene lactones: structural and biochemical aspects. Curr Org Chem 3(6):577–608
47. Seaman FC (1982) Sesquiterpene lactones as taxonomic characters in the Asteraceae. Bot Rev 48(2):121–594. https://doi.org/10.1007/BF02919190
48. Acevedo CH, Scotti L, Scotti MT (2018) In silico studies designed to select sesquiterpene lactones with potential antichagasic activity from an in-house asteraceae database. ChemMedChem 13(6):634–45
49. Pérez DJ, Zakai UI, Guo S, Guzei IA, Gómez-Sandoval Z, Razo-Hernández RS et al (2016) Synthesis and biological screening of silicon-containing ibuprofen derivatives: a study of their NF-κβ inhibitory activity, cytotoxicity, and their ability to bind IKKβ. Aust J Chem 69(6):662–71